A Novel Design for Estimating Relative Accuracy of Screening Tests When Complete Disease Verification Is Not Feasible

Todd A. Alonzo
Department of Biostatistics, University of Southern California Keck School of Medicine, 440 E. Huntington Drive, Suite 300, P.O. Box 60012, Arcadia, California 91066, U.S.A.
email: talonzo@childrensoncologygroup.org

and

John M. Kittelson
Department of Preventive Medicine and Biometrics, University of Colorado Health Sciences Center, 4200 East 9th Avenue B-119, Denver, Colorado 80262, U.S.A.
email: John.Kittelson@uchsc.edu

SUMMARY. The accuracy (sensitivity and specificity) of a new screening test can be compared with that of a standard test by applying both tests to a group of subjects in which disease status can be determined by a gold standard (GS) test. However, it is not always feasible to administer a GS test to all study subjects. For example, a study is planned to determine whether a new screening test for cervical cancer (“ThinPrep”) is better than the standard test (“Pap”), and in this setting it is not feasible (or ethical) to determine disease status by biopsy in order to identify women with and without disease for participation in a study. When determination of disease status is not possible for all study subjects, the relative accuracy of two screening tests can still be estimated by using a paired screen-positive (PSP) design in which all subjects receive both screening tests, but only have the GS test if one of the screening tests is positive. Unfortunately in the cervical cancer example, the PSP design is also infeasible because it is not technically possible to administer both the ThinPrep and Pap at the same time. In this article, we describe a randomized paired screen-positive (RPSP) design in which subjects are randomized to receive one of the two screening tests initially, and only receive the other screening test and GS if the first screening test is positive. We derive maximum likelihood estimators and confidence intervals for the relative accuracy of the two screening tests, and assess the small sample behavior of these estimators using simulation studies. Sample size formulae are derived and applied to the cervical cancer screening trial example, and the efficiency of the RPSP design is compared with other designs.

Key words: Cervical cancer; Inverse probability weighting; Pap; Screening; Sensitivity; Specificity; ThinPrep; Verification bias.

1. Introduction
Cervical cancer is one of the most common malignancies in women. For women in the United States, the lifetime risk of a cervical cancer diagnosis is 0.82%, and the risk of disease mortality is 0.29% (Reis et al., 2002). The incidence of cervical cancer and associated mortality have each decreased by over 40% in the last 30 years. This is sometimes attributed in large part to early diagnosis of precancerous lesions by screening women with the Papanicolaou (Pap) test (Schiffman et al., 1996). Despite the progress made, the Pap is not a perfect screening tool. The true positive fraction (TPF) or sensitivity is the proportion of women with cervical cancer who screen positive. The false positive fraction (FPF) or 1-specificity is the proportion of women without cervical cancer who screen positive. It is estimated that Pap has a TPF of around 50% and an FPF of about 2% (McCrory et al., 1999). The low TPF of the Pap has motivated the development of new screening tests in hopes that they will have better accuracy.

In an ideal study to compare the accuracy of a new screening test to the standard screening test, both screening tests and a gold standard (GS) test that provides a definitive determination of disease status would be administered to a random sample of subjects from the population of interest. In practice, however, it may not be ethical to employ this paired design because it requires the GS test which may be invasive (such as biopsy) to be administered to subjects who test negative on the new and standard screening tests. For example, it is difficult to justify recommending a cancer-confirming biopsy to women who screen negative both with the Pap and a new screening test because the results of the screening tests already suggest a low probability of cancer. Therefore, studies invariably only administer the GS to subjects who screen...
positive on at least one of the screening tests. Studies with this design are referred to as paired screen-positive (PSP) studies (Pepe and Alonzo, 2001). PSP studies have been used in the evaluation of screening tests for cervical cancer and in studies of other low prevalence diseases where the majority of study subjects screen negative on both screening tests.

In some settings, however, a PSP design may not be feasible. The example that motivates this work is a large randomized trial in New Zealand comparing the accuracy of a liquid-based cytology screening test called the “ThinPrep” with that of the Pap test. As described in Section 4, the choice between these two screening tests has important implications for the social equality and financial viability of the government-funded New Zealand cervical cancer screening program. In this setting a PSP design would require that all women receive both the ThinPrep and Pap at the screening visit. A woman would then receive colposcopy (the GS test) if one of the screening tests was positive. Both the ThinPrep and Pap require a sample of cells from the cervix (usually with a small brush, paddle, or swab), however it is not technically possible to use a single sample for both tests because the processing of the sample differs between tests. It is also not feasible to obtain two samples during the same visit because collection of the first sample often causes minor irritation of the cervix (e.g., bleeding), which would contaminate a second sample. A repeat screening test could be given after a 30-day wait, however compliance is likely to be very poor especially for women who screened negative on their initial test. The PSP design was not feasible for these reasons.

When there is concern that the implementation of an initial screening test could interfere with the implementation and results of a subsequent screening test, one could consider an unpaired screen-positive (USP) design where each subject receives one of the two screening tests and the GS only if the screening test is positive (Pepe, 2003, p. 183). Although such a design avoids the technical limitations of applying both screening tests, this design usually requires substantially more study subjects and more tests to be performed than the PSP design (Section 3.1).

In this article, we consider a study design in which subjects are randomized to receive one of the two screening tests. If that initial test is positive, then at a later time the subject receives the other screening test and the GS. We refer to this design as a randomized paired screen-positive (R PSP) design. This design can be thought of as a compromise between a PSP design and a USP design because with an RPSP design a subject may receive only one screening test (like a USP design) or both screening tests (like a PSP design) depending on the result of the first screening test.

In Section 2, we briefly review estimators of the relative accuracy and corresponding confidence intervals for PSP and USP designs, and derive estimators for an RPSP design. Sample size formulae provided in Section 3 are applied in Section 4 to a study comparing the accuracy of ThinPrep to Pap for detecting cervical cancer. We end with some recommendations and a discussion.

2. Estimators

If all subjects were to receive both screening tests plus the GS, then the accuracy of each screening test (TPF and FPF) would be estimable and provide a basis for comparison. A defining characteristic of screen-positive designs is that the GS test is only administered when the screening test is positive. Thus, TPF and FPF are not estimable because the disease status of subjects who screen negative is missing. Although absolute accuracy is not estimable, a screen-positive design yields estimates of the disease detection probability (DP; the proportion who screen positive and have the disease) and the false referral probability (FP; the proportion who screen positive and do not have the disease). Then, since the DP is equal to the product of TPF and disease prevalence, relative accuracy \( rTPF = \frac{TPF_A}{TPF_B} \) can be estimated by the ratio of detection probabilities \( rDP = \frac{DP_A}{DP_B} \):

\[
\frac{DP_A}{DP_B} = \frac{Pr(D+, A)}{Pr(D+, B)} = \frac{Pr(A | D+)}{Pr(B | D+)} Pr(D+) \]

where \( A+, B+, \) and \( D+ \) denote, respectively, a positive result on screen test A, screen test B, and the GS test. Similarly, FP is equal to 1 minus the prevalence of disease multiplied by FPF so that the relative false referral probability \( rFPF \) is equivalent to the relative FPF:

\[
\frac{FP_A}{FP_B} = \frac{Pr(D-, A)}{Pr(D-, B)} = \frac{Pr(A | D-)}{Pr(B | D-)} Pr(D-) \]

In the remainder of this section, we consider the problem of computing a point estimate and confidence interval for \( rTPF \). There are analogous procedures for \( rFPF \).

2.1 Paired Screen-Positive Design

Table 1 summarizes notation for a PSP design where all \( N \) study subjects are administered two screening tests (test A and test B) with binary results (+ or −) and the GS test that definitively determines binary disease status (D) is only ascertained if at least one of the screening tests is positive. Using the notation in Table 1, Schatzkin, Connor, and Taylor (1987) have shown that \( rTPF \) can be consistently estimated as

\[
\hat{rTPF}_{PSP} = \frac{\hat{Pr}(D+, A+) = (N_{DAB} + N_{DAB})/N}{\hat{Pr}(D+, B+) = (N_{DAB} + N_{DAB})/N} = \frac{N_{DAB} + N_{DAB}}{N_{DAB} + N_{DAB}} \]

where the numerator divided by \( N \) is an estimate of \( DP_A \) and the denominator divided by \( N \) is an estimate of \( DP_B \). Confidence intervals for \( rTPF \) have been based on approximate"
Table 2

Data from a USP study comparing the accuracy of screening tests A and B

<table>
<thead>
<tr>
<th></th>
<th>D+</th>
<th>D-</th>
<th>Total</th>
<th></th>
<th>D+</th>
<th>D-</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A+</td>
<td>$N_{DA}^A$</td>
<td>$N_{DA}^A$</td>
<td>$N_{DA}^A$</td>
<td>B+</td>
<td>$N_{DB}^B$</td>
<td>$N_{DB}^B$</td>
<td>$N_{DB}^B$</td>
</tr>
</tbody>
</table>

Table 3

Data from an RPSP study comparing the accuracy of screening tests A and B. In addition, the following data are known: the total number of study subjects ($N$), the number of subjects who are randomized to test A and test negative on test A ($N_{DA}^A$), and the number of subjects who are randomized to test B and test negative on test B ($N_{DB}^B$).

<table>
<thead>
<tr>
<th></th>
<th>D+, B+</th>
<th>D+, B-</th>
<th>Total</th>
<th></th>
<th>D-, B+</th>
<th>D-, B-</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A+</td>
<td>$N_{DAB}^A$</td>
<td>$N_{DAB}^A$</td>
<td>$N_{DAB}^A$</td>
<td>B+</td>
<td>$N_{DB}^B$</td>
<td>$N_{DB}^B$</td>
<td>$N_{DB}^B$</td>
</tr>
</tbody>
</table>

2.2 Unpaired Screen-Positive Design

Next, consider a USP design where $N^A$ subjects receive test A and $N^B$ subjects receive test B. Subjects who screen positive receive the GS test but do not receive the other screening test. The data from such a design are summarized in Table 2. The following estimator of rTPF has been previously proposed (Pepe, 2003, p. 184):

$$\hat{r}_{\text{TPF}_{\text{USP}}} = \frac{\Pr(D+, A+) \times N_{DA}^A / N^A}{\Pr(D+, B+) \times N_{DB}^B / N^B},$$

where $\Pr(D+, A+)$ is the proportion of subjects who screen positive on both tests A and B and have disease.

2.3 Randomized Paired Screen-Positive Design

Consider an RPSP design where $N$ subjects are randomized to first receive test A or test B. The second screening test and GS are only then administered if the first screening test is positive. Data from such a design are summarized in Table 3. For an RPSP design, superscripts are added to the notation of the PSP design to indicate the initial screening test subjects were randomized to receive first. For example, $N_{DAB}^A$ corresponds to the number of subjects who first screened positive with test A and subsequently tested positive on test B and were determined to be diseased by the GS test.

Using the notation in Table 3, equation (1) is equivalent to

$$\hat{r}_{\text{TPF}_{\text{PS}}} = \frac{N_{DAB} + N_{DAB}^B + N_{DAB}^A + N_{DAB}^B}{N_{DAB}^A + N_{DAB}^B + N_{DAB}^A + N_{DAB}^B},$$

This estimator of rTPF cannot be used with an RPSP design because unlike the PSP design, $N_{DAB}^A$ and $N_{DAB}^B$ are not observed. Therefore, we next derive an estimator of rTPF that is appropriate for an RPSP design.

Assume that the accuracy of the tests are not affected by the order in which the tests are performed (this is discussed further in Section 5) and that there are $N^A$ subjects who

$\text{Relative Accuracy of Screening Tests}$

Data from a USP study comparing the accuracy of screening tests A and B. In addition, the following data are known: the total number of study subjects ($N$), the number of subjects who are randomized to test A and test negative on test A ($N_{DA}^A$), and the number of subjects who are randomized to test B and test negative on test B ($N_{DB}^B$).

$\text{Relative Accuracy of Screening Tests}$
of subjects who test positive on both tests A and B. That is, \( \hat{\delta} \) is the ratio of the number of subjects who receive test A first and do not test positive on both test A and test B to the total number of subjects who do not test positive on both test A and test B.

The MLEs in (6)–(9) are Horvitz–Thompson inverse weighting estimators (Horvitz and Thompson, 1952) that weight \( N_{\text{DAB}}^A \) by \( 1/\hat{\delta} \) and \( N_{\text{DAB}}^B \) by \( 1/(1-\hat{\delta}) \) because \( N_{\text{DAB}}^B \) and \( N_{\text{DAB}}^A \), respectively, cannot be observed. To illustrate, consider the MLE for \( \delta \) (equation (6)). \( \widehat{\Delta}_A \) is the sum of estimators for \( \Pr(\text{D-}+, \text{A-}+, \text{B-}) \) and \( \Pr(\text{D}+, \text{A}+, \text{B}+) \). Since both screening test results and disease status are observed for subjects who test positive on tests A and B, \( \Pr(\text{D}+, \text{A}+, \text{B}+) \) is estimated as \((N_{\text{DAB}}^A + N_{\text{DAB}}^B)/N \) (equation (10)). Subjects who receive test B first and screen negative on test B do not receive test A and the GS so \( N_{\text{DAB}}^B \) cannot be observed. Thus, \( N_{\text{DAB}}^B \) is estimated by \( N_{\text{DAB}}^A (1-\hat{\delta})/\hat{\delta} \) (equivalently, \( N_{\text{DAB}}^A + N_{\text{DAB}}^B \) is estimated by \( N_{\text{DAB}}^A \hat{\delta} \)). Conceptually, \( \hat{\delta} \) is an estimator of the proportion of subjects who receive test A first. For example, with equal randomization we expect \( \hat{\delta} \approx 1/2 \), so \( N_{\text{DAB}}^A + N_{\text{DAB}}^B \approx 2N_{\text{DAB}}^A \), and the MLEs of equations (6)–(9) are the familiar multinomial estimates. Although it might seem more intuitive to use \( \hat{\delta} = N_A^A/N \) (the actual proportion who receive test A first), the estimator of (12) is required to assure that the estimated probabilities within each group sum to 1 (i.e., \((N_{\text{DAB}}^A + N_{\text{DAB}}^B + N_{\text{DAB}}^A + N_{\text{DAB}}^B)/N^A = 1 \) and \((N_{\text{DAB}}^A + N_{\text{DAB}}^B + N_{\text{DAB}}^A + N_{\text{DAB}}^B)/N^B = 1 \).

Since \( r_{\text{TPF}} = \widehat{\Delta}_A/\widehat{\Delta}_B \), the ratio of the MLEs for \( \widehat{\Delta}_A \) and \( \widehat{\Delta}_B \) yields the following MLE for \( r_{\text{TPF}} \):

\[
\widehat{r_{\text{TPF}}} = \frac{N_{\text{DAB}}^A + N_{\text{DAB}}^B + N_{\text{DAB}}^A/\hat{\delta}}{N_{\text{DAB}}^A + N_{\text{DAB}}^B + N_{\text{DAB}}^B/(1-\hat{\delta})},
\]

A 100 \((1-\alpha)\)% confidence interval for \( r_{\text{TPF}} \) is then given by

\[
\exp\left\{ \log \widehat{r_{\text{TPF}}} + Z_{(1-\alpha/2)} \sqrt{\text{Var}(\log \widehat{r_{\text{TPF}}})} \right\},
\]

where \( Z_\nu \) is the \( \nu \)-th quantile of the standard normal distribution and \( \text{Var}(\log \widehat{r_{\text{TPF}}}) \) is given in the technical report available at http://www.tibs.org/biometrics for the setting where there are \( N \) subjects per group.

Next, we derive an approximation to the likelihood-based variance. Consider the estimator obtained by using \( \hat{\delta} \) as opposed to \( \hat{\delta} \):

\[
\widehat{r_{\text{TPF}}} = \frac{\widehat{P_{\text{DAB}}}}{\widehat{P_{\text{DAB}}}/\hat{\delta} + \widehat{P_{\text{DAB}}}/(1-\hat{\delta})}.
\]

where \( \widehat{P_{\text{DAB}}} = (N_{\text{DAB}}^A + N_{\text{DAB}}^B)/N \), \( \widehat{P_{\text{DAB}}} = N_{\text{DAB}}^A/N^A \), and \( \widehat{P_{\text{DAB}}} = N_{\text{DAB}}^B/N^B \). These estimates will be approximately independent in settings with low disease prevalence because multinomial probabilities are effectively independent when the multinomial probabilities are small. Therefore, in large samples with low disease prevalence each of these estimates will be approximately independent and normally distributed:

\[
\widehat{P_{\text{DAB}}} \sim \mathcal{N}(P_{\text{DAB}}, P_{\text{DAB}}(1 - P_{\text{DAB}})/N) \quad \text{where} \quad P_{\text{DAB}} = \text{DDP};
\]

\[
\widehat{P_{\text{DAB}}} \sim \mathcal{N}(P_{\text{DAB}}, P_{\text{DAB}}(1 - P_{\text{DAB}})/N) \quad \text{where} \quad P_{\text{DAB}} = \text{DAP} - \text{DDP};
\]

\[
\widehat{P_{\text{DAB}}} \sim \mathcal{N}(P_{\text{DAB}}, P_{\text{DAB}}(1 - P_{\text{DAB}})/N) \quad \text{where} \quad P_{\text{DAP}} = \text{DAP} - \text{DDP}. \]

Using the delta method on (14) gives the following approximate variance:

\[
\text{Var}(\log \widehat{r_{\text{TPF}}}) = N^{-1} \left\{ \frac{1}{\delta P_{\text{A}}} (DP_A - DDP)(1 - DP_A + DDP) + \frac{1}{(1 - \delta) P_{\text{B}}} (DP_B - DDP)(1 - DP_B + DDP) + \left( \frac{1}{DP_A} - \frac{1}{DP_B} \right)^2 DDP(1 - DDP) \right\}.
\]

where \( \delta = N_A^A/N \). This variance can be estimated by substituting \( DDP_A \), \( DDP_B \), and \( DP_B \) with \( \widehat{P_{\text{DAB}}} \), the numerator of (14), and the denominator of (14), respectively.

In settings relevant to cervical cancer studies where disease prevalence is very low (1%), (15) and the likelihood-based variance differ by less than 0.3%. The difference between the two variance approximations increases as disease prevalence increases, however the difference remains small (<6.4%) even if disease prevalence is as high as 25%.

Relative to the likelihood-based variance, (15) has advantages that it is easier to calculate and it can be calculated as long as \( \widehat{\Delta}_A \) does not equal \( \widehat{\Delta}_B \), i.e., there is no perfect agreement between the tests, whereas the likelihood-based variance cannot be estimated if any of the observed cell counts in Table 3 are zero.

Simulation studies suggest that the estimators have good performance in small samples except in extreme settings when confidence intervals cannot be constructed (see technical report available at http://www.tibs.org/biometrics).

3. Sample Size Calculations

Since TPF quantifies the key benefit of screening and FPF conversely quantifies the key disadvantage of screening, it is important to consider both parameters when comparing the accuracies of screening tests. A new screening test would not be preferred over the standard test if, for example, the new test was superior in one of the parameters but the improvement came at the expense of substantial degradation in the other parameter. A new screening test (test A) may be preferred over the standard test (test B) if one of the following conditions is satisfied: (i) test A has superior TPF and FPF; (ii) test A has superior TPF and has FPF that is not substantially inferior; (iii) test A has superior FPF and has TPF that is not substantially inferior; (iv) both TPF and FPF for test A are not substantially inferior to test B and test A is substantially less invasive or costly than test B.

These four conditions can be tested using the composite null hypothesis \( H_0 \) : \( r_{\text{TPF}} \leq \gamma_1 \) or \( r_{\text{FPF}} \geq \gamma_2 \), where \( \gamma_1 = 1 \) tests whether test A has superior TPF (i.e., \( P_{\text{FA}} > P_{\text{FB}} \)) and \( \gamma_2 = 2 \) in some narrow range less than 1 tests whether test A has TPF that is not inferior to test B (i.e., \( P_{\text{FA}} > \gamma_1 P_{\text{FB}} \)) (Alonzo, Pepe, and Moskowitz, 2002). Defining the narrow range that constitutes noninferiority is often a difficult task that may depend on many factors including the costs of the tests, costs associated with false positives and false negatives, and the prevalence of disease in the population screened.

Similarly, \( \gamma_2 = 1 \) tests whether test A has superior FPF (i.e.,
FPF_A < FPF_B) and γ_2 in some narrow range greater than 1 tests whether test A has FPF that is not inferior to test B (i.e., FPF_A < γ_2 FPF_B).

The composite hypothesis can be tested using joint confidence limits for rTPF and rFPF. Specifically, the null hypothesis is rejected if the confidence interval for rTPF does not contain γ_1 and the confidence interval for rFPF does not contain γ_2. Therefore, condition (i) can be tested using γ_1 and γ_2 equal to 1, condition (ii) can be tested using γ_1 = 1 and γ_2 > 1, condition (iii) can be tested using γ_1 < 1 and γ_2 = 1, and condition (iv) can be tested using γ_1 < 1 and γ_2 > 1.

The power of the joint hypothesis test described above is the product of the power for each of its parts. Thus, if the study has power (1 − β_P) to detect superior rTPF and power (1 − β_F) to rule out inferiority, then the joint test has power (1 − β) = (1 − β_P)(1 − β_F). In general, the sample size could be chosen to control β_P, β_F, or β. With screen-positive designs, disease prevalence is usually low so that the sample size required for testing rTPF is usually much larger than that required for rFPF. In these situations β_F ≈ 0, so that the sample size required for the joint hypothesis is essentially the same as that required for a single hypothesis test of rTPF.

In this section, we consider the sample size requirements for testing rTPF. Sample size equations for estimating rFPF follow by analogy. Consider determining the sample size required to give power 1 − β against the alternative hypothesis rTPF = γ in a one-sided significance level-α test of rTPF = 1. Then, Alonzo et al. (2002) showed that the PSP design requires a total sample size of

\[ N_{\text{PSP}} = \left( \frac{Z_{(1-\beta)} + Z_{(1-\alpha)}}{\log \gamma} \right)^2 \frac{DP_A + DP_B - 2DDP}{DP_A DP_B}. \] (16)

With an equal number of subjects in each group, the variances in (4) and (15) yield that the USP and RPSP designs require total sample sizes of

\[ N_{\text{USP}} = 2 \left( \frac{Z_{(1-\beta)} + Z_{(1-\alpha)}}{\log \gamma} \right)^2 \frac{(1 - DP_A)(1 - DP_B)}{DP_A DP_B}, \] (17)

and

\[ N_{\text{RPSP}} = \left( \frac{Z_{(1-\beta)} + Z_{(1-\alpha)}}{\log \gamma} \right)^2 \times \left\{ \frac{2}{DP_A^2} (DP_A - DDP)(1 - DP_A + DDP) + \frac{2}{DP_B^2} (DP_B - DDP)(1 - DP_B + DDP) + \frac{1}{DP_A} (1 - DP_B)^2 DDP(1 - DDP) \right\}, \] (18)

respectively.

3.1 Comparison of Designs

Next, we compare the three screen-positive designs with respect to the number of subjects required, the number of GS tests required, and the total number of tests required. Table 4 gives the expected number of tests performed for the three screen-positive designs if there are N study subjects.

<table>
<thead>
<tr>
<th>Design</th>
<th>Test A</th>
<th>Test B</th>
<th>GS</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSP</td>
<td>N(1−Pr(B−))</td>
<td>N(1−Pr(A−))</td>
<td>N(1−Pr(A−)+Pr(B−))</td>
</tr>
<tr>
<td>RPSP</td>
<td>N(1−Pr(B−))</td>
<td>N(1−Pr(A−))</td>
<td>N(1−Pr(A−)+Pr(B−))</td>
</tr>
<tr>
<td>USP</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>

Pr(A−, B+ | D+) has been referred to as the true positive fraction (TPPF) and is equivalent to DDP divided by the prevalence. TPPF can be interpreted as a measurement of the amount of agreement (correlation) in the tests among diseased subjects. When there is no agreement in the screening test results for subjects with disease (i.e., TPPF or equivalently DDP equals zero), N_{RPSP} is equivalent to N_{USP} because by additionally testing those subjects who screen positive the RPSP design yields no additional diseased subjects. N_{RPSP} is twice N_{PSP} and N_{USP} when DDP = DP_A × DP_B which is approximately satisfied for studies with 1% disease prevalence when DDP equals zero. When there is high agreement, N_{RPSP} and N_{USP} are much smaller than N_{USP} because by additionally testing those subjects who screen positive the RPSP and PSP designs yield additional diseased subjects. As agreement increases from minimal to maximum, the difference between N_{RPSP} and N_{USP} decreases.

In the settings considered, the required number of GS tests for the PSP design is less than that for the RPSP design and significantly less than the USP design. The larger the amount of diseased agreement, the more similar the number of GS tests performed with the RPSP and PSP designs and the larger the discrepancy between these designs and the USP design. For no or very little diseased agreement, the three designs require a similar number of total tests to be performed. However, as the amount of agreement increases, the RPSP and PSP designs require significantly fewer tests.

As prevalence increases, the number of subjects required decreases, but the total number of tests required for RPSP and PSP compared with USP are similar for the 1–25% prevalence considered.

4. Cervical Cancer Screening Illustration

New Zealand has a government-funded national cervical cancer screening program (NCSP) for women aged 20–69 years. The program funds health promotion, smear taking, laboratory analysis of smears and biopsies, and management of women with abnormal smears. The NCSP uses the Pap test, but some doctors are offering women the ThinPrep for a small additional fee (~$15). If NCSP continues to fund only the Pap, then economically disadvantaged women would be less likely to opt for a ThinPrep test, which creates the potential for a socioeconomic inequality if the ThinPrep is in fact superior. On the other hand, the additional cost of the ThinPrep has important implications for the NCSP budget (Cox and
Sneyd, 2004). A randomized trial has been proposed in New Zealand to estimate the rTPF and rFPF of these tests and to determine whether the cost of the ThinPrep is warranted. The results of this trial will also inform similar national cervical cancer screening programs in Australia (Medical Services Advisory Committee, 2002) and the United Kingdom (Steering Group on the Feasibility of Introducing Liquid Based Cytology, 2002). As stated in Section 1, it is not possible in the New Zealand trial to perform both the Pap and ThinPrep at the same time; thus, RPSP and USP designs are feasible whereas PSP is not, even though it is included in the following discussion.

Consider designing a study to determine whether ThinPrep (Test A) has superior TPF and noninferior FPF compared with Pap (Test B). For the purposes of study design it is reasonable to base comparisons on the assumption that the prevalence of cervical cancer is 1% and TPF of Pap is 50% (i.e., DP_B = 0.005). The sample size required for power (1 - \( \beta \)) = 0.9 to detect rTPF = 1.3 (i.e., TPF of ThinPrep is 65%; DP_A = 0.0065) in a one-sided level \( \alpha = 0.025 \) test is given by equations (16)-(18), and varies by the amount of agreement between Pap and ThinPrep (Figure 1a). For example, if TPPF = 0.4 (i.e., DDP = 0.004; 80% of maximal agreement), then PSP and RPSP designs would require 16,439 and 31,515 women, respectively, whereas the USP design would require 107,417 women. If FPPF = Pr(A+, B+ | D-) = 0.04 (i.e., FFP = 0.04 \times 0.99) and FPF of Pap is 5% (i.e., FP_B = 0.05 \times 0.99), then if ThinPrep and Pap have the same

![Figure 1](attachment:figure1.png)

**Figure 1.** Sample size requirements and costs for testing the alternative hypothesis that the TPF of ThinPrep divided by the TPF of Pap (rTPF) is 0.65/0.5 = 1.3; USP (dashed line), RPSP (dotted line), and PSP (solid line). Prevalence of cervical cancer is 1%. TPPF = (TPF of Pap) \times \Pr(\text{ThinPrep}+ | \text{Pap}+) . (b)-(d) consider rFPF = 0.05/0.05 = 1 and FPPF = (FPF of Pap) \times \Pr(\text{ThinPrep}+ | \text{Pap}+) = 0.04.
specificity (rFPF = 1.0), an RPSP study with 31,515 women will have 98.8% power for rejecting a 10% increase in FPF (rFPF = 1.1). For the joint alternative hypothesis of rTPF = 1.3 and rFPF = 1.0, a joint null hypothesis test of rTPF ≤ 1.0 or rFPF ≥ 1.1 has power 0.9 × 0.988 = 0.89.

With more agreement there is not much difference between the number of GS tests or total number of tests performed with the PSP or RPSP design, and both outperform the USP design (Figure 1b and 1c). Study costs will vary by the number of the different types of tests according to their respective costs. In this trial the cost of a single Pap, ThinPrep, and colposcopy test is approximately $5, $20, and $200, respectively. Figure 1d compares the total costs of all tests in the three designs, and again shows clear advantages for the PSP and RPSP designs. For example, if TPPF equals 0.4, then the cost of PSP, RPSP, and USP designs are $630,924, $763,754, and $2,529,651, respectively. Since the PSP design is not feasible in this setting, the RPSP design appears to be the best choice.

5. Discussion

Verification bias (Begg and Greenes, 1983) occurs in studies where the selection of subjects to have disease status verification with a GS test is influenced by the results of the screening tests under evaluation. Since in screen-positive designs disease status is not determined unless at least one of the screening tests is positive, screen-positive studies have extreme verification bias for estimating absolute accuracy (TPF or FPF), and can only provide unbiased estimates of relative accuracy (rTPF or rFPF).

Screen-positive designs such as the PSP, USP, or RPSP will provide unbiased estimates of relative accuracy as long as disease verification depends only on the results of the screening test(s); that is, disease verification is missing at random (MAR; Little and Rubin, 1987, p. 12). Although screen-positive studies are designed so that disease verification is MAR, lack of compliance or lengthy delays between the screening and GS test can introduce bias. Specifically, if disease status changes in the interval between screening and GS tests, then relative accuracy estimates from the PSP, USP, or RPSP can be biased. With cervical cancer it is generally reasonable to assume that the screening test and colposcopy are measuring the same thing as long as colposcopy is given within 6 months of a positive screen. In general the risk of bias due to changes in disease status will be reduced if the time between the screening and GS tests is short, so trials using screen-positive designs should avoid excessive delays between tests. A related concern in the NZ example was that bias could result if: (a) both tests were given at the same visit (a PSP design), (b) the second sample was contaminated by the collection process for the first sample, and (c) if the interpretation of contaminated samples differed between tests. The RPSP design avoids this bias. Noncompliance (subjects who screen positive but do not receive disease verification) can also introduce bias, although an adjustment (Alonzo, 2005) could be considered if lack of compliance does not depend on disease status.

Other designs have been proposed for comparing the accuracies of two screening tests when the GS test is not administered to the subjects who screen negative. The partial testing design of Baker, Connor, and Kessler (1998) is a paired design where only a fraction of subjects negative on the first test get a more expensive second test and the GS is obtained by giving a test sometime later. The design in Baker and Pinsky (2001) differs from the PSP design discussed in this article in that only a fraction of subjects negative on the first test get the more expensive second test and the outcome measure is the ratio of partial receiver operating characteristic (ROC) curves rather than separate ratios of TPFs and FPFs. This outcome measure is not applicable when there are binary screening tests, as considered in this article.

The (low prevalence) approximate variance of (15) assumes independence between the estimated probabilities that are used for rTPF, which with multinomial data, is only true if these cell probabilities are small. We have found that this approximation gives good coverage probabilities in simulation studies with disease prevalence as high as 25%, and we conclude that it should perform well in screen-positive trials where prevalence would typically be well below that figure.

This article illustrates the RPSP design in the context of a trial in New Zealand to compare the ThinPrep and Pap tests for cervical cancer screening. This trial measures rTPF and rFPF by colposcopy as the GS. Although we have not described the details of how any of the three tests will be scored, the trial must, of course, adopt standard definitions to assure generalizability and reproducibility. In particular, criteria for a positive colposcopy must be carefully defined (ASCUS-LSIL Triage Study [ALTS] Group, 2003). We note that the RPSP design can be used to evaluate other types of cervical cancer screening tests (e.g., MonoPrep, Cytospin, SurePath), and in other rare disease settings with invasive GS tests when the screening tests cannot be given simultaneously. It is also important to note that the efficacy of cancer screening must ultimately be judged in trials with a mortality endpoint; thus, the designs described in this article apply to the early phases (Pepe et al., 2001) in the development of screening tests.

We conclude that an RPSP design can be used to obtain unbiased inference about the relative accuracy of two screening tests. Its efficiency is likely to be much better than a USP design, and may not be much worse than a PSP design. Thus, the RPSP design offers clear advantages when a PSP design is not feasible, and in some settings could be more convenient without much loss of efficiency even if a PSP design is feasible.

Acknowledgements

The authors thank Dr Margaret Pepe and an associate editor for helpful comments that led to an improved version of the article. This research is partially funded by NIH R01 GM54438. Computation for the simulation studies described in this article was supported in part by the University of Southern California Center for High Performance Computing and Communications (www.usc.edu/hpcc).

References


